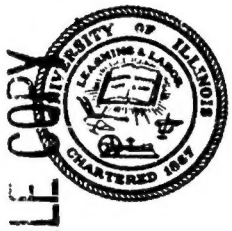
AD No. 25722
ASTIA FILE COPY



**BUREAU OF RESEARCH AND SERVICE**
**College of Education**
**University of Illinois**
**Urbana, Illinois**

# THE COMPARATIVE EFFICIENCY OF THREE TEST DESIGNS FOR MEASURING SIMILARITY BETWEEN PERSONS

Willard G. Warrington
Board of Examiners, Michigan State College

Study performed under Contract N6ori-07135
with the Office of Naval Research

Project on
Social Perception and Group Effectiveness

Technical Report No. 8
August, 1953

# THE COMPARATIVE EFFICIENCY OF THREE TEST DESIGNS FOR MEASURING SIMILARITY BETWEEN PERSONS[1]

Willard G. Warrington
Board of Examiners, Michigan State College

## The Problem

Techniques for studying the similarity between persons have been suggested repeatedly in psychological literature. The particular technique most frequently discussed involves the principle of "inverse" factor analysis or Q-technique as introduced by Burt and Stephenson in 1935 (1, 15). Stephenson (18) accords a great deal more importance to this approach than Burt (2) who sees little difference between "inverse" and conventional factor analysis.

The advent of World War II temporilarily diverted the attention of most psychologists, including Stephenson, from this area. Since 1946 however, there has been a great increase in the interest and application of Q-type correlational techniques. Researchers using Q-sort and Q-technique are having a certain amount of success in exploring complex problem-areas that in the past have been quite difficult to attack by more traditional methods. Many such studies could be listed but the following may be considered as a sample of the wide range of application of Q-technique: Fiedler's studies of patient-therapist relationships (8); Eberman's application of Q-technique to the measurement and prediction of teaching efficiency (6); Edelson and Jones' use of Q-technique in conjunction with role playing to investigate the self-concept (7); and Fiedler, Blaisdell, and Warrington's investigation of the relationship of social perception and unconscious attitudes (10).

These researchers who have had success with Q-technique have tended to turn their attention to new problems with little or no effort toward understanding and improving the technique itself. Until quite recently no adequate attempt has been made to examine the basic assumptions - and,

---

[1]This study was accomplished under Contract N6ori-07135 between the Office of Naval Research and the University of Illinois. This report is based in part on a doctoral dissertation submitted to the College of Education, University of Illinois (19).

particularly, the limitations underlying the Q-sort and Q-technique. Consequently this study is concerned with the following three difficulties:

(1) Lack of differentiation between the Q-sort and Q-technique. Q-sort is Stephenson's method of obtaining intrapersonal ratings by requiring the rater to sort adjectives or statements according to a specified distribution. Many investigators are apparently becoming confused in regarding the Q-sort as synonymous with Q-technique. They are ignoring the fact that the Q-sort was designed as a tool to obtain data that could be further analyzed according to the Q-technique rationale. The Q-sort was developed primarily as a convenient design for collecting data to be used in computing Q correlations, i.e., correlations between persons (16). The Q-technique however, is a much broader conceptualization. Its primary prerequisite is that persons be treated as variables, and traits or tests as populations. Certainly there is nothing in Stephenson's original development that suggests that the only data that can be treated in this manner are those that are obtained by the Q-sort.

Furthermore, there is considerable disagreement in the literature as to the value of forced-choice designs such as the Q-sort versus unforced designs. While most of these studies are not oriented toward measuring similarity between persons but rather toward accurately describing individuals, many of the issues are identical. For example, Zuckerman (20) found no apparent differences in the relative effectiveness of comparable interest inventories when unforced L-I-D responses were compared with forced-choice paired-comparisons. Gordon (11), however, found the forced-choice method to be more valid than the unforced questionnaire method in the measurement of four personality traits. Thus, it becomes necessary to examine the efficiency of the Q-sort as compared to alternative possibilities for studying the similarities between persons.

(2) Inadequate theory concerning the measurement of the similarity between persons. A recent paper by Cronbach and Gleser (5) is one of the first systematic attempts to examine the theoretical assumptions and limitations involved in several alternative methods of estimating similarities between persons. These writers point out possible serious effects of the forced distribution of responses that is required in the Q-sort. They discuss losses of information that may result when the means and variances of all individuals are equated as is required by this forcing.

Their paper has raised serious questions as to the advisability of using Q-sort designs as a means for obtaining similarities between persons. The investigation here being presented was conducted simultaneously with the Cronbach-Gleser study and the original design antedated their mathematical development. Since there has been considerable interaction between the two projects, a major purpose of this study will be to clarify and extend some of the implications suggested in their paper.

(3) Uncertainty as to relative merit of alternative Q-sort designs. As is generally true of any specific method, slightly different innovations have been developed by researchers using the Q-sort design in specific experimental situations.

An example of a novel variation of the Q-sort is reported by Fiedler, Hartmann, and Rudin (9). In an attempt to relate social perceptions to group effectiveness in basketball teams, these investigators used a personality questionnaire consisting of 100 items. These items were arranged in twenty blocks of five items each. The subject was asked to mark that item in each block that was most characteristic of him and also that item in each block that was least characteristic of him. In such a design, each block could be considered as a miniature Q-sort with one item (the least characteristic) going in the 0 pile, three items (the neutrals) going in the 1 pile, and one item (the most characteristic) going in the 2 pile. The responses to this block design were treated as in the Stephenson Q-sort; Q correlations were obtained and various significance tests were applied.

This paper reported several interesting and provocative conclusions. However, one could ask whether this is really a Q-sort design. Would the obtained results have been ditferent if the regular Q-sort procedure had been followed? How does the efficiency of this block-type instrument compare with the efficiency of the Q-sort or with other alternative methods for investigating the similarities between persons? The present state of knowledge concerning Q-sort designs is simply not sufficient to permit clear answers to these and many other questions.

## The Methodology

There are two methods of studying the efficiency of a particular test design; (1) a rational or mathematical analysis, or (2) an empirical approach using real or artificial test situations. In the mathematical analysis the investigator attempts to design a mathematical model that approximates the theoretical structure of the test situation. Then by studying the effect of variations of elements in the model he makes predictions as to the efficiency of the test. Such an approach is by far the more powerful since mathematical analysis of limiting conditions permits generalization as to the maximum range of effect. Furthermore, over-all trends and potential pitfalls are more readily pointed up by the mathematical approach.

In a relatively unexplored and unfamiliar area, however, there may not be sufficient understanding of the underlying structure to permit the identification of relevant variables. It is not possible, therefore, to pursue a mathematical analysis in any efficient manner. In such a situation the empirical approach can often be used to give more adequate direction to the mathematical attack. This is particularly true when specific and practical questions are to be answered, for the empirical approach may provide information that will enable the investigator to ask better questions of the mathematical analysis.

Too often an imperfect tool must be used simply because our state of knowledge is not sufficient to permit the construction of a better instrument. In such a situation it is important to determine the loss in efficiency that results from the use of this poorer instrument. This type of information is usually difficult or impossible to obtain except by the empirical approach.

## Plan of Attack of this Study

In view of the inadequate knowledge concerning the efficiencies of different test designs for measuring the similarity between persons, the present study follows an empirical approach using artificial data.

In broad outline the program of this study proceeds in the following order:

1. Hypothetical data, defined mathematically, are used throughout the study.

2. An adequate criterion is structured and justified.

3. Similarity measures or scores are computed for various test designs.

4. These obtained similarities are compared with the criterion similarities to determine the efficiency of the various test designs.

This procedure is followed using a perfectly reliable test, i.e., an error-free case, and repeated with a test of a moderate degree of reliability, i.e., with error added. This permits an analysis of the efficiency of a test design in terms of both validity and reliability.

## Specific Purposes of this Study

The purposes of this study are essentially as follows:

1. To develop a theoretical model for investigating the efficiency of various types of instruments for measuring the similarities between persons.[2]

2. To apply this model in investigating the efficiency of the Q-sort methods for measuring similarities between persons as compared to an alternative method that does not require forcing. In particular, losses of information due to the forced distribution of responses in the Q-sort will be investigated.

3. To use this model to compare the efficiency of two different types of Q-sorts, a total Q-sort and a block Q-sort. Specifically,

---

[2] The rationale, mathematical structure, and computational requirements of this theoretical model will not be discussed in this paper. However, a complete and detailed description is available in the doctoral dissertation of the writer (19). A microfilm copy of this dissertation is available at the University of Illinois Library or will be supplied to Navy agencies or contractors by this project.

a Q-sort, similar to that developed by Stephenson, will be compared with a modified block Q-sort design, similar to that used by Fiedler, Hartmann, and Rudin (9).

4. To investigate the effect of error in measuring similarities between persons. Each of the test designs are studied under error-free conditions and again after a specified amount of error has been introduced into the model.

5. To study the effects of cluster scoring Q-sort responses, i.e., combining scores of homogeneous items to obtain cluster scores. Results obtained by scoring all items separately and in clusters are compared as to validity and reliability.

These purposes are somewhat extensive and range from the exploratory to the relatively specific. Therefore, many of the questions suggested by the accomplishment of these purposes can, at best, be answered only partially. However, it is hoped that some of these questions can be answered completely, and that several new but meaningful questions can be raised for further investigation.

## Application of the Model

The procedure for studying the efficiency of a test as developed in this study rests on the following rationale. A test is primarily used for predictive purposes, that is, the tester is interested in predicting the score or standing on some specified criterion from the obtained results on the test. A particular test would be considered efficient if the test results consistently reproduce these criterion scores. This criterion may be such a variable as success or failure on a job, staying out of certain institutions or staying in others, or often, when nothing better is available, the score on another test or even a total score on this particular test.

The theoretical model that has been developed to investigate the efficiency of various test designs for measuring the similarity between persons permits the specification of "true" criterion measures. Consequently the relative efficiencies of various test designs can be compared under ideal or "perfect" testing conditions.

While the detailed structure of the theoretical model cannot be discussed here the application of the model to any particular test design involves the following basic steps.

1. The criterion domain is determined by specifying the number of orthogonal factors or dimensions in the test space.

2. The hypothetical subjects are assigned "true" scores on these dimensions thereby fixing their position in the test space.

3. The distances[3] between these persons are computed and constitute the true criterion similarity scores.

4. The test items are located in the variate space as a function of the factorial loadings and popularity of the item.

5. The "obtained" score of each person on each item is computed by mathematical manipulation.

6. These obtained scores are used to assign item scores according to conditions imposed by the specific test design.

7. Obtained similarity scores are computed from the item scores and compared with the true criterion similarity scores to give efficiency estimates.

The Specific Conditions and Assumptions of this Study

Due to the delimited area of operation of this phase of the study, the foremost criteria in defining the structure of this experimental model are those of parsimony, reasonableness, and usefulness. An attempt is made to approximate actual test situations that are being reported in the literature with as simple a model as seems logically reasonable.

In accordance with these criteria, the model under consideration is structured according to the following specifications.

1. The factor or test space consists of five orthogonal factors.

2. N, the number of persons in the variate space is twenty. Since our criterion space consists of the distances between these twenty persons this space has an $N^1$ of 190 such distances. This number of persons is then quite adequate for our purposes.

3. The population of persons, of which the specified twenty are a sample, is normally distributed over each of the five factors.

4. Furthermore, all factors are assumed to have the same variance over the population of persons. Conditions 3 and 4 are rather restrictive but seem more suitable in this exploratory stage than more complex assumptions.

5. As a consequence of 3 and 4, each factor distribution can be transformed to a standard normal distribution. This transformed variate space has its origin at ( 0, 0, 0, 0, 0,) with unit variance on each factor.

---

[3] These distances are computed by the formula for the D measure as discussed by Cronbach and Gleser (5) and Osgood and Suci (13).

Assignment of factor scores. The five conditions just listed are sufficient to permit the assignment of factor scores to each of the twenty persons. These factor scores can be treated as coordinates and thus fix the persons' positions in the space. The equality of factor variances permits the specification of an underlying metric. Since each factor distribution has been standardized this metric is the unit variance. Each person can thus be assigned a score on each factor by randomly drawing standard scores from a standard normal table.

The factorial structure of the hypothetical test items. Having fixed the position of each of the twenty persons in the test space, the conditions necessary to construct the hypothetical test items are now considered.

6. The hypothetical test consists of sixty items. This number is somewhat arbitrary since Q-sorts have been reported with considerably more than sixty items. A sixty-item test, however, is convenient to use for the conditions of this study and seems sufficiently long to insure adequate sampling in practice if the items are properly structured.

7. Each item has non-zero loadings on at least two and no more than five factors. That is, univocal items are not treated. This seems reasonable since relevant items loaded only on a single factor are very difficult to write.

8. Each item has a loading of .70 or .80 on one of the five factors. Such items are more possible to write and this structure permits cluster scoring, that is, obtaining a factor score by adding the scores of all items heavily loaded on that particular factor. The variation of .70 and .80 is suggested as a means of reducing ties in the obtained scores. Such ties may sometimes be difficult to manipulate under the conditions of the Q-sort.

9. The heavily loaded items specified in condition 8 are equally distributed over all factors. Specifically, each of the five factors has six items with loadings of .70 on that factor and six items with loadings of .80 on that same factor. Therefore, this test consists of five distinct clusters of twelve items each.

10. The loading of the first factor is always positive or zero. If the first factor loading is zero then the second factor is positive, etc. This restriction delimits the type of items here being considered to those that have a meaningful positive direction.

11. Those factor loadings unspecified by conditions 8, 9, and 10 are assigned randomly as to magnitude and sign but subject to the constraint that the sum of the squares of all loadings must total unity for each item. Since the criterion space has been completely specified as consisting of five factors, no item is permitted to have loadings on a factor not included in this space.

12. Item popularities (percent of population that would accept or endorse the item) are normally distributed within the total test and equally distributed within each cluster. These popularities range from ten to ninety percent.

The above twelve conditions are sufficient to establish the criterion similarity scores between the 190 pairs of persons. Furthermore, the sixty test items have been specified so that obtained "item" scores can now be computed for each person. The actual item score of each person on each item is determined by the nature of the test design into which that item is incorporated.

## The Three Experimental Test Designs

The same sixty hypothetical items are used throughout this study. However, obtained similarity scores are computed under three different designs. Obviously, no attempt has been made to examine the characteristics of all, or even a considerable portion, of the possible types of test designs that might be used to study similarities between persons. Furthermore, since the underlying elements of the theoretical model are held constant over the entire empirical investigation, variations in efficiency for any of the test designs that would result from changes in factor structure, test structure, or any specified assumption or condition are not evident from this presentation.

It is shown, however, that trends are apparent in the comparisons of different test designs between and within different criterion spaces. The criteria for selection and specification of these test designs were based on simplicity, reasonableness, and usefulness. These trends, therefore, should have significance for the investigator who is actually involved in the study of similarity between persons at either the theoretical or operational level.

The total Q-sort. This forced-choice design approximates the original Q-sort as described by Stephenson. The entire item sample is sorted by each subject in the specified distribution indicated in Table 1 and the similarities between persons are computed from these Q scores.

TABLE 1

DISTRIBUTION OF RESPONSES FOR TOTAL Q-SORT

| Item Score for each Category | Theoretical Number of Items in Each Category | Number of Items Assigned to Each Category |
|---|---|---|
| 0 | 0.5 | 1 |
| 1 | 3.3 | 3 |
| 2 | 9.8 | 10 |
| 3 | 16.4 | 16 |
| 4 | 16.4 | 16 |
| 5 | 9.8 | 10 |
| 6 | 3.3 | 3 |
| 7 | 0.5 | 1 |
| Totals | 60.0 | 60 |

The block Q-sort. As stated earlier an intent of this study is to investigate the efficiency of the block-type Q-sort described by Fiedler, Hartmann, and Rudin (9). This design involves a modified Q-sort in which the total item sample is subdivided into blocks of five items each. The subjects are asked to select the most positive item from each block according to different sets of instructions. The resulting scores are treated as if the conventional Q-sort design had been followed.

The unforced design. The third test design that is treated has little in common with the Q-sort method. This test is completely unforced in that the response of each individual on any item is permitted to take any value - theoretically, from plus infinity to minus infinity.

In practice, no such design could be operationally structured since such precise measurements do not exist in the social sciences. A more reasonable test and one that could approximate the requirements of non-forced and continuous scores might be designed so that the response to each item could be indicated on a multi-position scale. For example, the Likert-type scale (3) would involve items somewhat similar to those of the unforced test considered in this study. The ideal test was selected to serve as a sort of referent test against which all other alternative structures could be compared.

## Error

The efficiencies of these three test designs for the conditions specified above are computed under two conditions; first, under the condition of perfect reliability, i.e., the error-free case, and second, after a constant rate of error has been introduced in the test items.

Since this study is of exploratory nature, rather restrictive assumptions are made concerning this introduction of error. This seems justified in that this study is more interested in the general effect of error rather than specific reliabilities. Error is assigned under these assumptions.

1. Error is normally distributed over all factors.

2. The rate of error will be assumed equal for all factors.

3. The rate of error will be assumed equal for all persons.

Since computational complexities demanded that a single rate of error be introduced it was decided that this error rate should represent a moderate reliability (approximately .80) since most tests used in similarity studies do not have higher reliabilities.

# THE RESULTS

## Properties of Individual Profiles

Prior to the discussion of the results of this study, some attention must be given to the techniques of profile analysis since the basic data are in the form of profiles. For each person in this study, a score has been assigned on each of five factors. These scores then represent the true profile of the individual. Likewise, other profiles consisting of item scores assigned under varying test conditions have been computed for each individual. These item profiles represent estimates of his true profile. The criterion similarity scores between persons are determined by computed distances between true profiles. Obtained similarities are computed between the comparable test profiles. The efficiency of a given test design is estimated by studying the relationship between the criterion similarities and obtained similarities for both the error and error-free cases.

Cronbach and Gleser point out that differences between two individual profiles have three components. There may be differences in elevation, in scatter, and in profile shape. These terms are defined as follows. Elevation is the mean of all the scores for a specific person. Scatter is the square root of the sum of squares of the individual's deviation scores about his own mean or elevation. Shape is the information left in the profile after differences due to elevation and scatter are removed.

It is possible, therefore, to study similarities between persons under various conditions. For instance, by converting all profile scores to deviation scores about the profile mean, similarities can be computed that are independent of differences in elevation. Similarly, by standardization within each profile, i.e., converting all raw profile scores to standard scores with mean of zero and standard deviation (s.d.) of unity, the effects of differences in elevation and scatter can be eliminated. In such a standardizing process the only information left within each profile is that due to shape. A specific example is presented to illustrate the possible transformations.

Person 2 has a profile of -1.5, -0.5, -0.5, 0.5, 0.0 with mean of -0.4 and s.d. of .66.

Person 7 has a profile of 1.5, 2.5, -1.0, 0.5, -3.0 with mean of 0.1 and s.d. of 1.93. Then, by the formula suggested by Cronbach and Gleser (5, p. 4).

$$D_{27}^2 = (-1.5 - 1.5)^2 + (-0.5 - 2.5)^2 + (-0.5 + 1.0)^2 + (0.5 - 0.5)^2 + (0.0 + 3.0)^2$$

$$= 27.25, \text{ and}$$

$$D_{27} = 5.22 .$$

Differences in elevation can be eliminated by converting the raw score profiles to deviate scores. These deviation profiles are:

Person 2: -1.1, -0.1, -0.1, 0.9, 0.4

Person 7: 1.4, 2.4, -1.1, 0.4, -3.1

Now $D_{27}$ with differences in elevation ruled out is 5.10. Since the elevation scores of these two persons are relatively similar in size these two distances are also similar.

Differences in elevation and scatter can be eliminated by converting the profile score sets of both persons to standard score profiles with means of zero and s.d. of unity. The standardized profiles are:

Person 2: -1.7, -0.2, -0.2, 1.4, 0.6

Person 7: 0.7, 1.2, -0.6, 0.2 -1.6

$D_{27}$ under the conditions of equal scatter and equal elevation is 3.76. This trend from 5.22 to 5.10 to 3.76 is logical since persons 2 and 7 differed in elevation, scatter, and shape. Eliminating successive components reduced the dissimilarity and brought them closer together in the variate space.

These results are consistent with the mathematical model developed by Cronbach and Gleser. They described the process of eliminating differences in elevation as a projection of all persons into a variate space of one less dimension. This is accomplished by the passing of a hyperplane[4] through the variate space perpendicular to the elevation factor and projecting all persons onto this hyperplane.

Within this hyperplane another projection is possible that equates the scatter of all individuals. This is accomplished by again projecting the position of all persons onto a hypersphere of one less dimension. In a three-dimensional structure, elevation effects are removed by projecting all persons into a two-dimensional plane. Within this plane all persons are further projected into a circle with the center at the

[4]The terms, hyperplane and hypersphere, refer to the generalized plane and sphere concepts in a space of more than three dimensions.

origin and the radius determined by the common variance that is being specified for all individuals. [For further explanation of the mathematical system for reducing profiles, see the Cronbach-Gleser report (5)] .

## Criterion Similarity Scores

### Development of Criterion Measures

In view of the above discussion three different types of criterion similarities can be computed. Persons can be compared as to similarity between profiles when all information is to be treated. Persons can also be compared for similarity when differences in elevation are not considered or when differences in both elevation and scatter are eliminated.

Since the test space has five orthogonal factors these three criterion measures have 5, 4, and 3 dimensions or degrees of freedom, respectively. To simplify notation the criterion similarity score-set in which all the information is treated, i.e., the similarity score that operates with five degrees of freedom, is designated as $C_5$. Similarly, the score-set in which the dimension of elevation is not treated is $C_4$ and the criterion score-set in which differences in elevation and scatter are eliminated is $C_3$. All three of these configurations of similarity scores involve the same 190 interrelationships between the twenty persons but different information will be included in each type of score.

### Comparison of the Three Criterion Score-sets

The intercorrelations[5] between the three criterion score-sets are as follows: $C_5$ and $C_4$ have a correlation of .82, $C_5$ and $C_3$ of .39, and $C_4$ and $C_3$ of .49. The correlation of .82 between the five-and-four space measures suggests that the operation of eliminating differences in elevation is quite costly. The $C_5$ measures include all the information that accounts for the criterion variance. Reducing the criterion space one dimension by treating similarities between deviation profiles as in $C_4$ loses 33 percent[6] of the total criterion variance.

This loss of true criterion information is even more pronounced when differences in scatter are also ruled out as in $C_3$. This is accomplished by standardizing the deviation profiles thereby specifying the same profile variance of unity for all persons. The correlation between the true criterion $C_5$ scores and these reduced $C_3$ scores is .39. This

---

[5] These intercorrelations are product moment correlations. The distributions of the D measures were slightly skewed but not to the extent that product moment correlations were appreciably effected.

[6] The square of the correlation between the $C_5$ and $C_4$ scores indicates the percentage of the $C_5$ variance that is explained by the variance of the $C_4$ scores (12, p. 116).

correlation indicates that the variation in the three-dimensional space, consisting of D measures representing only shape differences, represents approximately 15 percent of the variance in the total criterion space.

It would be possible, under the conditions of this study, for a test instrument to produce measures that would correlate highly with the $C_3$ score-set and yet have very low validity when the unreduced score-set is the desired criterion. Similarly, the $C_3$ scores contain only 24 percent of the information retained in the $C_4$ scores.

These data suggest that for maximum reproduction of the total criterion score-set, similarity measures should be used that treat all the information in the variate space. Operationally, however, this does not mean that similarity scores of k - 1[7] or possibly even k - 2 dimensions must never be used. If the elevation dimension is clearly irrelevant to the similarity under investigation it would be poor judgment to include this information just to increase the dimensions of the measuring space. The decision as to relevancy of the elevation score is sometimes difficult to make. For example, the elevation score may, in some tests, be determined primarily by a positive response set, i.e., the subject tends to reply positively more often than negative. This will create an artificial common factor that may or may not be relevant in the similarity under consideration. Many researchers would hold that such response sets are irrelevant and should be treated as error variance or when possible eliminated. On the other hand, it is possible to argue that this positive response set is an indication of an optimistic attitude and as such, is very important in certain similarity measures.

It is much more difficult to build a case for eliminating scatter as well as elevation. The fact that some studies have reported positive results using test designs in which these variables have not been considered is not evidence of the value of this procedure. In view of the huge loss of information that our data indicate occurs when scatter is equalized, one would suspect that in spite of the positive results reported other more important differences were probably overlooked due to the inefficiency of the approach.

The implications of these extreme losses of information must be interpreted with some caution. One must keep in mind that this study has specified that each of the twenty persons can have scores only on the factors in the $C_5$ space, i.e., it does not permit specific factor scores.

---

[7] When referring to the general variate space of k dimensions, the k - 1 similarity score will be the similarity score with differences in elevation eliminated. Similarly, the k - 2 similarity score has differences due to both elevation and scatter eliminated.

It is difficult to predict how far one could generalize, if, for instance, each person had a score on one or more of five factors plus a score on a factor unique to him. Certainly the magnitude of effects in this 25-space would be reduced but the same trends should be present. The effects under more complex conditions will need further study.

## Overview of the Total Data of this Study

In addition to the three different criterion similarity score-sets discussed above, two obtained similarity score-sets, one without error and one with error, were computed for each comparable score for each of the three test designs. To allow for more adequate comparison of these configurations of similarity scores Table 2 lists these various score-sets and briefly summarizes the properties of each.

It will be recalled that each score-set consists of 190 similarity measures. Each of these similarity measures is computed using the Cronbach-Gleser D formula. The D measures that constitute the $DQ_3$ score-set for the total Q-sort and the block Q-sort are computed directly from Q correlations.

In the computation of the $DQ_3$ measures each subject's profile consists of the scores on the 60 items. The mean and variance of these profiles are held constant over all persons by the forcing requirements of the Q-sort design. Each subject, therefore, responds to the Q-sort items under two restrictive conditions and, as a consequence, loses two degrees of freedom in his selection of item responses. Since the total Q-sort contained 60 items, each response distribution has 58 degrees of freedom. Similarly, in the block Q-sort design the 60 items are divided into 12 blocks of five items in each block. Each block represents a small Q-sort. Therefore, a subject loses two degrees of freedom in responding to each block and, hence, loses 24 degrees of freedom in responding to the entire 60-item test. His responses, therefore, carry 36 experimentally independent judgments.

It must be kept in mind, however, that the entire item vector space is defined to have five orthogonal dimensions, i.e., a five-factor space. The 58 degrees of freedom or dimensions of the total $DQ_3$ similarity scores and the 36 dimensions of the block $DQ_3$ scores represent subspaces that lie within the complete five-variate space. Thus, the specificed dimensions of the $DQ_3$ scores of the total and block Q-sorts can not all be orthogonal dimensions. This is not unexpected since the

Table 2

SUMMARY OF SIMILARITY SCORE-SETS

| Type of Similarity Score-Set | Symbol | No. of Degrees of Freedom | Similarity Measures Computed from | Individual Differences that Contribute to Similarity Measures | Individual Differences not Considered in Similarity Measures |
|---|---|---|---|---|---|
| Criterion | $C_5$ | 5 | true factor profiles | elevation, scatter, shape | All information treated |
| | $C_4$ | 4 | deviation factor profiles | scatter, shape | elevation |
| | $C_3$ | 3 | standardized factor profiles | shape | elevation, scatter |
| Unforced | $D_5$ | 5 | obtained cluster profiles | elevation, scatter shape | all information treated |
| | $D_4$ | 4 | deviation cluster profiles | scatter, shape | elevation |
| | $D_3$ | 3 | standardized cluster profiles | shape | elevation, scatter |
| Total Q-sort | $DQ_3$ | 58* | Q correlations | shape | elevation, scatter |
| | $DQ_4$ | 4 | clustered Q score profiles | scatter, shape | elevation |
| | $D_3$ | 3 | standardized clustered Q score profiles | shape | elevation, scatter |
| Block Q-sort | $DQ_3$ | 36* | Q correlations | shape | elevation, scatter |
| | $DQ_4$ | 4 | clustered Q score profiles | scatter, shape | elevation |
| | $D_3$ | 3 | standardized clustered Q score profiles | shape | elevation, scatter |

* These measures are computed from unclustered item Q scores. See text for further explanation of the number of degrees of freedom.

item sample has been structured to contain items that cluster about each of the five factors. The items within any cluster, therefore, have high intercorrelation and tend to fall along one of the five a priori specified factor dimensions. Consequently, even though a subject has 58 degrees of freedom or independent judgments in marking the items in the total Q-sort, his item responses, nevertheless, all lie within the five-factor space.

Since the $DQ_4$ similarity measures are computed from cluster profiles, each profile consists of five cluster scores. Furthermore, the sum of these five cluster scores is equal for each subject since the sum of the 60 item scores is constant for all persons. Thus, the $DQ_4$ scores are computed under one constraint, and one degree of freedom is lost. These $DQ_4$ similarity distances, therefore, are vectors lying in a space of four dimensions. The $D_4$ and $C_4$ similarity scores for the unforced test and criterion configurations also represent vectors in a four-dimensional space. These D measures represent distances between deviation profiles, i.e., profiles with a constant mean of zero. At this point, however, it should not be assumed that these $D_4$ and $DQ_4$ obtained measures contain the same information, i.e., operate in the same four-dimensional space. This point is discussed in more detail later.

Similarly, those D measures that are computed between standardized profiles operate in k-2 dimensions - in this case, three dimensions - since standardized profiles have constant means and variances. Again it cannot be said without further examination that the $D_3$ similarity scores for the Q-sort contain information regarding the identical three dimensions as that contained in the unforced $D_3$ and $C_3$ score-sets.

## The Efficiency of the Unforced Test

The data relating to the efficiency of the unforced test design will now be discussed. The variations in efficiency as error is introduced into the system will be of particular interest. The advantages and limitations of the unforced design for measuring similarities between persons will be considered. Recommendations as to the advisability of using the unforced test to obtain similarity scores will be developed.

### Test-Criterion Relationships

The correlations between the criterion score-sets and the obtained similarity score-sets for the unforced test design are presented in Table 3. There are two comparable sets of correlations, one for the perfectly reliable test and the other for the moderately reliable test.

TABLE 3

CORRELATIONS OF CRITERION SCORE-SETS AND OBTAINED SIMILARITY SCORE-SETS FOR THE UNFORCED TEST DESIGN*

| Obtained Score-Sets | Perfectly Reliable Items | | | Moderately Reliable Items | | |
|---|---|---|---|---|---|---|
| | Criterion Score-Sets | | | Criterion Score-Sets | | |
| | $C_5$ | $C_4$ | $C_3$ | $C_5$ | $C_4$ | $C_3$ |
| $D_5$ | .92 | .58 | .27 | .81 | .55 | .25 |
| $D_4$ | .85 | .93 | .46 | .71 | .77 | .31 |
| $D_3$ | .39 | .44 | .90 | .26 | .31 | .45 |

*The correlation between each criterion score-set and its logically related, i.e., most relevant, obtained score-set is underlined. Correlations that represent semi-relevant relationships between criterion and obtained score-sets are underlined with dots.

The proper interpretation of these correlations regarding the efficiency of the unforced test requires some explanation since the wide range of values, from .93 to .27 for the error-free case, may be somewhat confusing. Consider, for example, the correlation between the $C_3$ and $D_5$ score-sets for the error-free test. How should this relationship of .27 be interpreted? Does this low correlation reflect a high or low efficiency for the unforced test?

The characteristics of the $C_3$ and $D_5$ scores have been discussed but, to review, a $D_5$ measure treats all the information in the unforced cluster profiles while a $C_3$ measure does not treat true-score information regarding differences in elevation and scatter. In other words, a $C_3$ measure and a $D_5$ measure are operating in different dimensions and treating different information. As a result, the $D_5$ distribution of scores should not correlate highly with the $C_3$ distribution since it will be recalled that the $C_5$ score-set had a correlation with $C_3$ of only .39.

It is necessary, therefore, to be quite explicit as to what is meant by the term efficiency. In this study a test has been defined as efficient if it consistently reproduces a relevant criterion. But what is a relevant criterion? Would the $C_3$ score-set be a relevant criterion for the unforced $D_5$ obtained score-set? The answer is no since the $D_5$ scores contain some information that has been deliberately eliminated in the $C_3$ scores. The correlations that determine the efficiency of the test design are those that describe relationships between obtained and criterion score-sets that are based on the same information, i.e., computed in comparable dimensions. Certain low criterion-obtained correlations may give information as to test efficiency, however, since a test score-set to be efficient should have a low correlation with a criterion score-set that it should not reproduce on logical grounds.

A test design can be examined for efficiency in reproducing the three different criterion score-sets - - $C_5$, $C_4$, and $C_3$. No single obtained score-set however, will be able to reproduce all three criterion similarity configurations with high accuracy since each criterion score-set contains different information.

The correlations that compare criterion-obtained score-sets that are computed in comparable dimensions and, hence, are most meaningful in determining the efficiency of the unforced test are those that lie along the principal diagonal in each correlation matrix. These correlations are underlined in Table 3 and are the largest of the criterion correlations of each obtained score-set as would be expected. The correlations of the $C_5$ and $D_4$ score-sets are partially underlined since it will be shown that the obtained information due to elevation that is eliminated in the $D_4$ score-set is different from the criterion information due to elevation that is eliminated in the $C_4$ score-set. This effect results from a specific condition of this study and, as a consequence, the $D_4$ scores include some of the criterion information due to the elevation component and, hence, is a semi-relevant obtained score-set for the $C_5$ score set.

For the error-free test the most relevant criterion-obtained correlations are .92, .93, and .90. These high correlations indicate that the perfectly reliable unforced test is quite efficient and also flexible. This flexibility is indicated by the fact that any of the three criterion score-sets can be reproduced to a high degree by a corresponding obtained score-set of the same dimensions.

The generally low correlations between criterion and obtained score-sets of dissimilar dimensions are further indication of the high efficiency of specific test score-sets for reproducing specific criterion information. This wide range of correlation between the various criterion and obtained score-sets suggests the importance of specifying

the measurement domain of similarity studies. In any similarity study one should, first, decide what dimensions are important to measure and, second, make sure that the obtained similarity measures include only those dimensions. Otherwise, much of the information contained in the measured similarity scores may be due to differences on dimensions that are irrelevant to the criterion information. For example, if elevation is important in the criterion then the obtained similarity measure must include elevation differences to be most valid. Similarly, if the researcher concludes on logical or empirical grounds that certain dimensions, such as scatter and elevation, are not relevant to the criterion, then these dimensions should not be included in the obtained measure.

## The Effect of Error on the Efficiency of the Unforced Test

Efficiency relationships must never be interpreted solely from error-free data, for consistency (i.e., reliability) is a necessary condition for high efficiency. Comparison of the criterion-obtained correlations for the perfectly reliable test with those of the less reliable test for the unforced case indicates a reduction of the relationship in every case. This is to be expected since the random error that has been introduced in the less reliable obtained configurations should have no relation to the criterion configurations, which are not affected by error. Since each reported correlation is between an obtained score-set and a criterion score-set, those relationships involving obtained similarity scores under error conditions should be smaller than comparable non-error correlations.

The criterion-obtained correlations between score-sets of comparable dimensions for the error data remain the highest in each row but are no longer of equal magnitude. The correlations $C_5$-$D_5$[8] and $C_4$-$D_4$ remain fairly high and approximately equal at .81 and .77, respectively, but the $C_3$-$D_3$ relationship drops off decidedly to .45. In interpreting such changes in patterns of correlation under non-error and error conditions, it must be kept in mind that a single rate of error was introduced into the theoretical model. If there are no process differences in the treatment of error, the pattern of relationships should be approximately constant under either the absence or presence of error. Since the pattern of correlations has changed as well as the magnitude of relationships in the unforced design, it must be concluded that error is treated differently by the three obtained similarity measures. In the unforced test the effect of error has been considerably magnified in going from deviation cluster profiles to standardized profiles, i.e., in the operation involving the equating of individual scatter. This magnification of error was expected since Cronbach and Gleser discuss this effect in considerable detail in their paper (5).

---

[8] For simplicity, the correlation between two similarity score configurations will be denoted by their respective symbols separated by a dash, e.g., the correlation of $C_5$ and $D_5$ will be written as $C_5$-$D_5$. This must not be interpreted as the difference between the two score-sets.

## More Empirical Evidence of the Effects of Error on the Unforced Test Design

The effect of error can be investigated by a direct empirical approach if two sets of obtained similarity measures have been computed for the same rate of error. The correlation of these two error score-sets would be comparable to a test-retest reliability estimate. Such a reliability coefficient could be computed for each of the three obtained similarity score-sets, i.e., the $D_5$, $D_4$, and $D_3$ configurations. Since these reliabilities were computed for a constant rate of error, any variation in the magnitudes of correlations would indicate differences in the effect of this error within the respective obtained similarity spaces.

In the present study, however, only one set of obtained scores was computed for the specified rate of error. While this condition eliminated the possibility of obtaining a test-retest reliability, it does permit the computation of a correlation that has a somewhat similar interpretation. Each obtained similarity score-set for the non-error case can be correlated with the comparable similarity score-set for the error case. All three obtained score-sets for the less reliable items should contain approximately constant proportional amounts of error if the rate of error was unchanged during the manipulation of profile information. Hence, there should be relatively constant correlations between error and non-error measures within each obtained space. If error is treated differently in the various obtained score-sets, these correlations will not be of constant magnitude and any consistent variation will reflect operational differences in the treatment of error.

The correlations between error-free similarity scores and the comparable scores containing error have been computed for each similarity configuration obtained under the unforced test design. These relationships are as follows; .85 for the $D_5$ scores, .83 for the $D_4$ scores, and .48 for the $D_3$ scores. These correlations are consistent with the conclusions of the theoretical discussion since they, also, indicate that the effect of error for the unforced test is practically the same for both the $D_5$ and $D_4$ similarity measures. Furthermore, the $D_3$ relationship, being much smaller, indicates that the rate of error has been considered magnified during the further reduction of the profile information.

## The Unforced Test and Similarity Measurement

The perfectly reliable unforced test as specified in this study has high efficiency for measuring the similarity between persons. This type of test is sufficiently flexible that the three different criterion score-sets can all be reproduced to a high degree by the corresponding obtained similarity score-sets. As error is introduced into the system the test efficiencies are decreased. For a moderate rate of error, however, the $D_5$ and $D_4$ efficiencies remain quite high, but the efficiency of the $D_3$ measures decreases decidedly as error is introduced.

Under practical test conditions, the unforced test design as here described should be adequately efficient in measuring the similarity between persons in k and k-1 dimensions. Similarity scores in k-2 dimensions should be obtained with considerable caution, but with items of moderate reliability the unforced design should be fairly adequate for these measures, also.

Limitations of this unforced test design. The unforced design as here described has at least two potential limitations that might affect its efficiency in measuring the similarity between persons. The first limitation concerns the difficulty of collecting data under the conditions specified for this unforced test. The second concerns the effect of response-sets upon similarity measures obtained under the unforced design.

The unforced test under consideration requires unforced continuous scores. To completely meet these requirements in practice each item would have to be accompanied by a continuous scale ranging from highly positive through the zero or neutral position to highly negative. Each subject would have to be able to specify his position accurately on this scale. The obtained item score would be determined by measuring the distance of the person's position from the zero position and assigning this distance the proper sign.

Operationally, such requirements are unrealistic, indeed, they are virtually impossible to fulfill. Subjects have definite limitations as to their ability to make extremely fine discriminations. The closest conventional approximation to the continuous unforced design would be represented by items that were to be marked on a five or seven point scale. These scale intervals should range from "most characteristic" or "most like me" through the neutral position to "least characteristic" or "least like me."

Since this broad interval classification is different from the continuous scores that were specified in this study, the results here presented must be interpreted with caution. While patterns of relationships should remain approximately constant, the magnitudes of these relationships may vary with actual data. This is particularly important to keep in mind when the unforced test efficiencies are compared with the efficiencies of the total and block O-sorts.

The second limitation of the unforced design presented in this study and, in fact, the greatest single disadvantage of the unforced design, in general, is that response-sets may affect the scores of many individuals. These response-sets may tend to reduce relevant individual differences and introduce variations on irrelevant factors. This may result in lowered discrimination on those variables that are considered important.

The effect of response-sets on unforced designs has been one of the strongest arguments favoring the forced-choice type of test. Since the

response distribution is specified in advance and held constant over all subjects, there is little opportunity for individual response-sets to affect the total test score. As a consequence, the forced-choice design has been described as potentially of high validity in many areas where response-sets make most unforced tests inadequate (4).

The argument that the unforced test design permits response-sets to appear in the test scores must not be taken lightly. Maximum effort should always be made to present clear and precise test instructions. If a multi-position scale is used, it is extremely important that each scale position be sharply defined. This is necessary in order that all subjects will react to the scale from a common frame of reference. Careful test design and item selection, thus, should tend to reduce the potential negative effect of response-sets on scores obtained from unforced tests.

General conclusions concerning the use of the unforced test in similarity studies. This study indicates that the unforced test design, as here described and tested with hypothetical data, can be quite efficient for obtaining similarity scores under moderate rates of error. Furthermore, this type of test is sufficiently flexible that different dimensions of criterion information can be efficiently measured by statistically eliminating the irrelevant dimensions from the obtained similarity measures. When error is introduced the efficiency of the k-2 obtained scores declines. Hence, the reliability of the items should be investigated prior to the interpretation of k-2 relationships.

The unforced test items should be so designed that each subject can give unforced and continuous responses to each item. While this can never be completely accomplished in practice, items with five or seven well-defined scale positions should approximate the ideal structure. Further research is needed to substantiate this inference since unforced items may operate less efficiently in practice than is indicated by the data from this hypothetical design.

It is usually possible to reduce the harmful effects of irrelevant response-sets in unforced items by careful structuring of the test instructions and the scale design on which responses are to be given. In some instances, relevant response-set effects can be incorporated into the similarity scores as valid variance.

## The Efficiency of the Total Q-Sort Design

The total Q-sort design considered here requires that all subjects assign each of the sixty test items a score from 0 to 7 according to a specified constant distribution. The data relating to the efficiency of this test design under the conditions specified in this study will now be presented. Information losses due to the forced design will be discussed. The effect of cluster scoring will be considered. The effect of a range of popularity of items within a Q-sort will be discussed. Recommendations as to the advisability of using the total Q-sort for measuring the similarity between persons will be developed.

## Test-Criterion Relationships

The correlations between the criterion and obtained similarity score-sets for the total Q-sort are presented in Table 4. The error-free correlations will be considered first.

TABLE 4

CORRELATIONS OF CRITERION SCORE-SETS AND OBTAINED SIMILARITY SCORE-SETS FOR THE TOTAL SIXTY-ITEM Q-SORT*

| Obtained Similarity Score-Sets | Perfectly Reliable Items | | | Moderately Reliable Items | | |
|---|---|---|---|---|---|---|
| | Criterion Score-Sets | | | Criterion Score-Sets | | |
| | $C_5$ | $C_4$ | $C_3$ | $C_5$ | $C_4$ | $C_3$ |
| $DQ_3$** | .78 | .70 | .66 | .22 | .18 | .38 |
| $DQ_4$ | .68 | .74 | .71 | .65 | .66 | .38 |
| $D_3$ | .36 | .40 | .88 | .24 | .28 | .38 |

*The correlation between each criterion score-set and its logically related, i.e., most relevant, obtained score-set is underlined. Correlations that represent semi-relevant relationships between criterion and obtained score sets due to specific conditions of this study are underlined with dots.

** The D measures for these configurations were computed from Q correlations.

The highest correlation for each criterion score-set lies along the principal diagonal of the error-free matrix. In this matrix, these largest relationships vary somewhat in magnitude. These data indicate that the $D_3$ score-set reproduces the $C_3$ criterion score-set more highly than the $DQ_4$ scores reproduce the $C_4$ scores or the $DQ_3$ scores reproduce the $C_5$ score-set. Specifically, the $C_3$-$D_3$ correlation of .88 indicates

that for perfectly reliable items the $D_3$ scores reproduce 77 percent of the information in the $C_3$ score-set. Similarly, the $DQ_4$ scores reproduce 55 percent of the $C_4$ information and the $DQ_3$ scores reproduce 61 percent of the $C_5$ information.

It is difficult, however, to properly interpret the correlations of the total Q-sort in reference to the efficiency of this particular test design. An obtained score-set has been described as efficient if it reproduces the information contained in the criterion score-set of comparable dimensions. Conversely, an efficient score-set should not have high relationships with a criterion score-set that includes information not contained in the obtained scores. These data, however, indicate that for perfectly reliable items the $DQ_3$ and $DQ_4$ scores reproduce approximately the same percentages of the criterion information regardless of what criterion score-set is being considered.

It will be recalled that the criterion measures have been defined so that the $C_5$ scores include all criterion information; the $C_4$ scores include all information except individual differences in profile elevation; and the $C_3$ scores include information due to individual differences in profile shape with elevation and scatter differences eliminated. Furthermore, the intercorrelations of the different criterion configurations, as listed on page 13 indicate that considerable criterion information is lost when differences due to elevation, or elevation and scatter, are not treated.

Since the $DQ_3$ similarity scores are computed between individuals' profiles that have constant means and variances over all persons, these scores do not contain information due to individual differences in obtained profile elevation or scatter. Likewise, the $DQ_4$ similarity scores are computed between obtained score profiles that have equal means over all persons, hence, differences in elevation are eliminated. Yet these $DQ_3$ and $DQ_4$ scores have approximately the same correlation with all three criterion score-sets.

Since a score-set should most highly reproduce that criterion score-set that includes the same information, these $DQ_3$ and $DQ_4$ scores apparently reproduce some criterion information due to individual differences in all three criterion profile components, namely, elevation, scatter, and shape. This suggests that the information components due to elevation and scatter are different for the criterion profiles than for the obtained score profiles. This suggestion will be examined before further consideration of the correlations between the criterion score-sets and the obtained similarity score-sets for the total Q-sort.

## Comparison of the Elevation Component of Criterion Profile Scores and Obtained Profile Scores

It will be helpful in better understanding these and other criterion-obtained score relationships if the term, elevation, is more clearly described. Profile elevation is defined as the mean of a profile of scores. The sum of the scores in the profile is a comparable variable since division of the sum by the number of scores in the profile produces the mean.

Different types of profiles are considered in computing the criterion and obtained similarity score-sets. Does the elevation component represent the same information in both types of profiles? This question must be answered before adequate interpretations of the various criterion-obtained correlations are possible.

Consider the true profile of an individual, i.e., his true scores on the five factors. Elevation for this type of profile is designated as the simple sum of the five factor scores in the profile. Now consider the elevation component of a profile consisting of the sixty obtained scores of a person for the item sample under consideration. What is the elevation component in this type of profile?

It will be necessary to introduce some new notation to demonstrate the elevation component of this obtained profile.

Let $i$ = items from 1 to 60,

$j$ = persons from 1 to 20,

$k$ = factors from 1 to 5,

then $x_{kj}$ = the true score of person j on factor k,

$\lambda_{ik}$ = the factorial loading of item i on factor k,

$X_{ji}$ = the obtained score of person j on item i, and

$SV_i$ = the distance of item i from the origin (i.e., the scale value of item i is a function of its popularity).

The obtained score of person j on item i has been defined to be (see above),

$$X_{ji} = \sum_{k=1}^{5} \lambda_{ik} x_{kj} - SV_i \ .$$

Therefore, the sum of the scores of person j on all sixty items (i.e., elevation) is,

$$\sum_{i=1}^{60} X_{ji} = \sum_{i=1}^{60} \sum_{k=1}^{5} \lambda_{ik} x_{kj} - \sum_{i=1}^{60} SV_i \ .$$

But $\sum_{i=1}^{60} SV_i = 0$ since popularity values have been balanced around the fifty percent position. Thus,

$$\sum_{i=1}^{60} X_{ji} = \sum_{i=1}^{60} \sum_{k=1}^{5} \lambda_{ik} x_{kj} = \sum_{k=1}^{5} x_{kj} \sum_{i=1}^{60} \lambda_{ik} \ .$$

Set $\sum_{i=1}^{60} \lambda_{ik} = w_k$, where $w_k$ is the sum of the loadings of all items on factor k, i.e., $w_k$ represents the weight of factor k in the total item sample. Since the item sample is constant over all persons, the obtained elevation component of person j is, therefore,

$$\sum_{i=1}^{60} X_{ji} = \sum_{k=1}^{5} w_k x_{kj} \ .$$

This equation indicates that the elevation component of a person's obtained profile is a function of the total item loadings on each factor, i.e., the factor weights, and the true score of the person on each factor. Furthermore, since the sixty-item test used in this study is composed of groups of homogeneous items that cluster about each of the five factors, this elevation component is the same for the obtained sixty-item profile as for the obtained cluster score profile.

Using the same notation, the elevation component in the total criterion profile for person j is $\sum_{k=1}^{5} x_{kj}$ .

A comparison of these two relationships indicates that the criterion and obtained elevation components will be identical, if and only if, the factor weights of the item sample correspond exactly with the factor weights in the criterion space.

In this study the criteric factor weightings and item sample factor weightings differ somewhat due to the requirement of the positive loading on $f_1$ (see page 7). Consequently, the information that is removed by eliminating elevation differences in the obtained profiles is not identical with the information that is removed by eliminating elevation differences in the criterion profiles.

A simple geometric example in two dimensions is presented in Figure 1 to illustrate the differences in these two elevation components. The criterion space is represented by two equally weighted orthogonal factors $f_1$ and $f_2$. If $x_{kj}$ is the score of person j on $f_k$, then the direction of the criterion elevation component in this space is represented by the line $L_1$ which has the equation $x_{1j} = x_{2j}$. This criterion elevation component can be eliminated by projecting all persons into a k - 1 space in which all persons have the same elevation. In the diagram, this k - 1 space is the line $L_2$ which has the equation $x_{1j} + x_{2j} = 0$.

Now suppose that the item sample selected to measure this criterion information has a disproportionately high factor loading on $f_2$; i.e., the factor weight $w_2$ is greater than the factor weight $w_1$. The direction of the obtained elevation component for such an item sample might be represented by the line $L_3$ which has the equation $w_1x_{1j} = w_2x_{2j}$. As in the criterion space, the line $L_4$ represents the k - 1 space into which the persons' obtained scores are projected to eliminate the elevation component. The equation of $L_4$ is $w_1x_{1j} + w_2x_{2j} = 0$. Since $w_1 \neq w_2$, the line $L_4$ is oblique to $L_2$.

Two persons, $p_1$ and $p_2$, are located in the total criterion space. Person $p_1$ has a greater mean (i.e., elevation) score than $p_2$ in the total space. In the criterion k - 1 space, i.e., after the elevation component has been removed by projection onto the line $L_2$, person $p_1$ and $p_2$ are identical since they project into the same position $p'$. The effect of removing elevation in the obtained score profiles, i.e., projecting all persons onto $L_4$ is considerably different. Person $p_1$ projects to $p_1'$ and $p_2$ projects to $p_2'$. The distance $d_{p_1', p_2'}$ represents the similarity of persons $p_1$ and $p_2$ in the obtained k - 1 space. Furthermore, the distance $d_{p_1', p_2'}$ has a component, $\triangle L_1$, parallel to the line $L_1$. This indicates that even though the obtained elevation component has been eliminated in the obtained scores, it is still possible for some of the criterion elevation component to remain in these obtained k - 1 profiles.

Generalizing this example to a five-dimensional space comparable to the variate space of this study suggests these implications. The location of the k - 1 or four-dimensional hyperplane onto which all persons are projected to compute their obtained k - 1 scores is dependent, in part, upon the factor loadings of the specific item sample. The criterion scores were computed under the specification of equal factor weights while the obtained scores were computed from an item sample of unequal

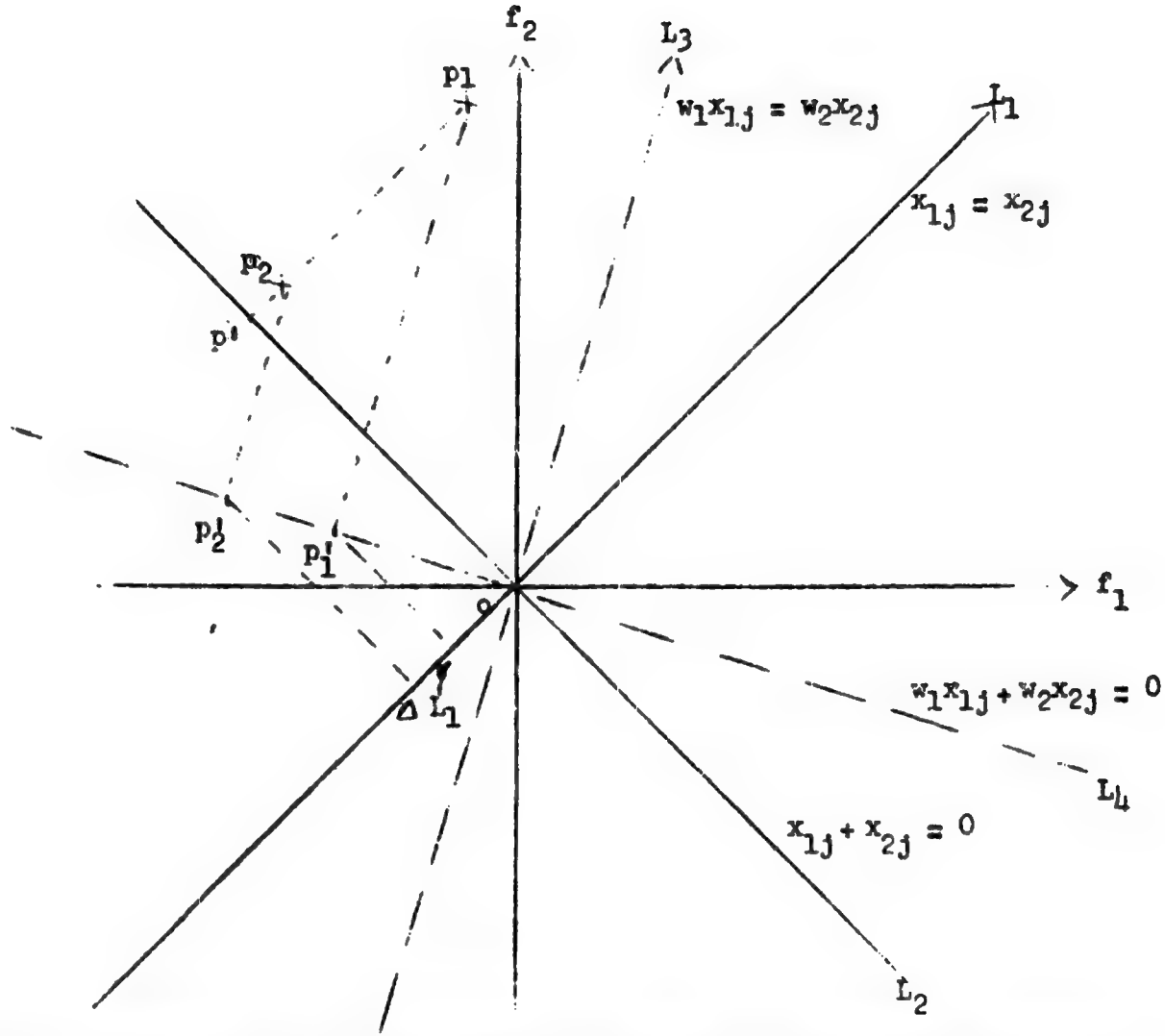



**Figure 1. Comparison of the Elevation Component of Criterion and Obtained Similarity Profile Scores.**

factor weightings. Therefore, the elevation component of the obtained item scores in this study is not identical with the elevation component of the criterion scores.

By similar reasoning, the reduction to k - 2 dimensions by eliminating differences in scatter is not an identical process for both criterion scores and obtained scores. Considering Figure 1 again, the k - 2 projection for the criterion scores would be accomplished in $L_2$ while the comparable projection for the obtained item scores would occur in $L_4$. Generalizing to a five-dimensional system, the k - 2 projection would be made in two different k - 1 hyperplanes if the criterion and item sample have different factor weights. Therefore, the information due to individual differences in profile scatter may vary as to its factorial content for obtained and criterion scores.

In practice, it is impossible to construct an item sample that weights all factors precisely as they are weighted in the selected criterion structure. Thus, in any actual item sample some of the differences between the criterion and obtained components of elevation and scatter, as described above, would exist from chance alone. In this study, however, the specification of a positive first factor loading forced the pattern of factor weights of the item sample to be considerably different from the pattern of factor weights of the criterion space.

Until further research is accomplished under more general conditions, the conclusions and implications of these data must be generalized with considerable caution since different results might be obtained if the items are randomly selected without any constraints. The above discussion, however, reinforces the suggestion made earlier as to the importance of understanding the structure of the criterion similarity relationships that are to be estimated by the obtained similarity scores. For not only is it necessary to know what criterion information is relevant but it is also necessary to know the structure of the criterion so that an item sample can be designed that will measure this relevant information.

With this background discussion, the correlations between the criterion and obtained similarity score-sets for the total Q-sort will be examined further. Since the same item sample has been used throughout this study, the comparison of different test designs should reflect differences primarily due to the characteristics of these test designs rather than specific characteristics of the item sample. Consequently, a comparison of the criterion-obtained relationships of the unforced and total Q-sort should give indications of the information lost due to the forcing requirement of the Q-sort.

## Information Lost Due to the Forcing Requirement of the Total Q-Sort

The amount of information lost due to the forcing requirements of the total Q-sort should be indicated by the comparison of the correlations

between the obtained score-sets and the most relevant criterion score-sets for the unforced and total Q-sort designs. The criterion that is most relevant to a particular obtained score-set contains the same kind of information as the obtained score-set within the limits determined by the specific characteristics of the item sample. For example, the $D_5$ scores of the unforced test are most relevant to the $C_5$ scores since neither score-set is computed under any constraint except those imposed by the specificity of the conditions of this study. The $D_4$ and the $DQ_4$ score-sets are computed from profiles that have constant means over all persons and, hence, will be most relevant to the $C_4$ score-set if the elevation components of the criterion and obtained profiles are approximately equal. As has just been shown, however, the specific conditions of this study cause these two elevation components to differ somewhat. As a result, the $D_4$ and $DQ_4$ score-sets, also, have some logical relationship with the $C_5$ score-set. Similarly, the $DQ_3$ scores and the $D_3$ scor for both the unforced and total Q-sort designs are computed between profiles that have constant means and variances over all persons. Thus, these scores are most relevant to the $C_3$ score-set but have some logical relationship with the $C_5$ score-set due to the unequal factor weights of the item sample as discussed above. It will also be shown in a later portion of this report that the $DQ_3$ scores include some information due to criterion scatter; hence, the $DQ_3$ score-set has some logical relationship with the $C_4$ score-set.

The relationships between the criterion score-sets and the most relevant obtained score-sets of the unforced and total Q-sort tests are presented in Table 5, for both error-free and error conditions. The correlations between the criterion score-sets and those obtained score-sets that are logically related because of specific conditions of this study[9] (i.e., semi-relevant) are also listed, but in parentheses. In every case, the unforced score-set reproduces more of the information in the related criterion score-set than the comparable score-set for the total Q-sort. This trend is true under the influence of error as well as for the error-free case. The data of Table 5 suggest that the unforced scores retain more relevant criterion information than the similarity scores of the total Q-sort. This loss of information is presumably due, in part, to forcing in the total Q-sort.

---

[9] Obviously, the differences between .90 and .88 or between .45 and .38 may be due to chance effects. It is extremely unlikely, however, that all differences could be attributed to chance.

TABLE 5

COMPARISON OF THE CORRELATIONS* OF CRITERION SCORE-SETS AND LOGICALLY RELEVANT OBTAINED SCORE-SETS FOR THE UNFORCED AND TOTAL Q-SORT DESIGNS

| Perfectly Reliable Items | | Moderately Reliable Items | |
|---|---|---|---|
| Criterion vs Unforced | Criterion vs Total Q-sort | Criterion vs Unforced | Criterion vs Total Q-sort |
| $C_5$-$D_5$ = .92 | ($C_5$-$DQ_3$ = .78) | $C_5$-$D_5$ = .81 | ($C_5$-$DQ_3$ = .22) |
| ($C_5$-$D_4$ = .85) | ($C_5$-$DQ_4$ = .68) | ($C_5$-$D_4$ = .71) | ($C_5$-$DQ_4$ = .65) |
| $C_4$-$D_4$ = .93 | $C_4$-$DQ_4$ = .74 | $C_4$-$D_4$ = .77 | $C_4$-$DQ_4$ = .66 |
| | ($C_4$-$DQ_3$ = .70) | | ($C_4$-$DQ_3$ = .19) |
| $C_3$-$D_3$ = .90 | $C_3$-$D_3$ = .88 | $C_3$-$D_3$ = .45 | $C_3$-$D_3$ = .38 |
| | $C_3$-$DQ_3$ = .66 | | $C_3$-$DQ_3$ = .38 |

*The relationships between criterion score-sets and those obtained score-sets that are logically related (semi-relevant) because of the specific conditions of this study are listed in parentheses.

It will be shown, that under certain conditions, the amount of this lost information can be reduced by cluster scoring, i.e., combining the scores on homogeneous items to obtain cluster scores. However, the total amount of information lost by forcing is impossible to determine since it has been shown that the magnitudes and patterns of criterion-obtained correlations are influenced by specific characteristics of the item sample. The characteristics of the different score-sets of the total Q-sort will now be considered further.

Information included in the total Q-sort $DQ_3$ scores. The criterion correlations involving the error-free $DQ_3$ similarity scores will be considered first. These correlations as listed in Table 4 are $C_5$-$DQ_3$ = .78, $C_4$-$DQ_3$ = .70, $C_3$-$DQ_3$ = .66. Since these correlations decrease as individual differences in elevation and scatter are eliminated from the

criterion scores, it must be concluded that the $DQ_3$ scores contain some information pertaining to criterion differences in elevation and scatter. At first glance, a consideration of the nature of the $DQ_3$ scores fails to confirm this conclusion.

The $DQ_3$ similarity scores have been computed from Q correlations. These Q correlations represent the relationship between one person's Q scores on the sixty items and the comparable sixty Q scores of a second person. The forcing requirement of the Q-sort requires that all persons have the same mean and the same variance. Thus, individual differences in elevation and scatter have been eliminated from the obtained similarity scores computed between these sixty score profiles. Since both $C_3$ and $DQ_3$ scores have eliminated information due to differences in elevation and scatter, how can $C_5$-$DQ_3$ be greater than $C_3$-$DQ_3$?

These relationships illustrate the effect of this item sample very clearly. While both the $C_3$ and $DQ_3$ scores have eliminated the elevation and scatter dimensions, these scores have not necessarily eliminated the same information. As has been demonstrated, the criterion factor weights differ from the factor weights of the item sample, hence, the components of elevation and scatter are different in criterion and obtained scores.

The $DQ_3$ scores, thus, contain some information due to criterion differences in elevation and scatter because of the specific characteristics of the item sample used in this study. As the structure of the factorial loadings of the item sample approaches the factorial loadings of the criterion space, the $DQ_3$ scores would become more similar to the $C_3$ scores. Under such conditions, the $C_3$-$DQ_3$ correlation should be greater than either the $C_4$-$DQ_3$ or $C_5$-$DQ_3$ correlations.

The introduction of random error into the system should have a slight tendency to reduce the effect of unequal factor weighting in the item sample since this error should be equally weighted for all factors. The $DQ_3$ correlations under the influence of error are $C_5$-$DQ_3$ = .22, $C_4$-$DQ_3$ = .18, and $C_3$-$DQ_3$ = .38. These data support the above inference since the $DQ_3$ scores correlate more highly with the $C_3$ scores than with the $C_5$ scores for the less reliable items.

The most obvious implications of these $DQ_3$-criterion correlations, however, are not concerned with patterns of relationships but rather with the magnitudes of the decreases in these correlations as error is added. The greatly reduced correlations under the influence of error indicate that the effect of the imposed error was magnified by the forcing requirement of the $DQ_3$ scores. Thus, these data indicate that the amount of information lost in the $DQ_3$ scores due to forcing increases decidedly as the test items become less reliable.

Further research will be necessary to determine the losses of information due to forcing under other test situations. Also, an investigation using several rates of error would be valuable to more adequately describe the effect of error on the $DQ_3$ similarity scores. This error effect was considered in this study but to a limited degree since only two rates of error were considered. Moreover, one of these was for zero error, i.e., the perfectly reliable case.

Information included in the $DQ_4$ scores and the effect of cluster scoring for the total Q-sort. The characteristics of the $DQ_4$ scores will reflect the advantages or disadvantages of cluster scoring. Briefly, the steps leading to the computation of a $DQ_4$ score were as follows:

1. The profile composed of Q scores on sixty items for each individual was transformed into a profile composed of five cluster scores. Each cluster score consisted of the sum of the scores of the twelve items that had loadings of .70 or higher on a particular factor. Thus, each cluster score represented the individual's obtained score on one of the five factors that define the test space.

2. The $DQ_4$ similarity score was computed using the Cronbach-Gleser D measure for obtaining the similarity between pairs of these five-score cluster profiles.

Since all the item score profiles had identical means, the cluster profile means were also equal for all persons. The scatter or variance within each item profile was constant over all persons, but after clustering the scatter within the individual cluster profiles differed over persons. Cluster scoring had apparently made available certain individual differences in profile scatter that were not treated in the profiles consisting of the total Q-sort item scores.

It remains to be shown that there is a correspondence between the scatter within these cluster profiles and the scatter within criterion profiles. If the individual differences in cluster profile scatter have little or no relationship to individual differences in criterion profile scatter, then this cluster scatter is irrelevant information. On the other hand, suppose there is a relationship between the scatter within these cluster profiles and the scatter within the individual criterion profiles. This would indicate that cluster scoring tends to reproduce some of the individual differences in scatter that are included in the criterion scoreset.

It is possible to examine this relationship directly by actually comparing the individual scatter within each criterion profile with the individual scatter within each obtained cluster profile. Since it was not

possible to claim any a priori knowledge as to the distribution of these scatters within the individual criterion profiles and obtained cluster profiles, the method of rank order correlation was followed.

The ranks of the scatters within each criterion profile were correlated with the ranks of the scatters within each obtained cluster profile for the twenty persons for each test design under both presence and absence of error. The correlations are presented in Table 6. The error-free correlations indicate the magnitude of this relationship under ideal conditions. The correlations under the influence of error indicate the

TABLE 6

RELATIONSHIP OF INDIVIDUAL CRITERION PROFILE SCATTER AND INDIVIDUAL OBTAINED PROFILE SCATTER AFTER CLUSTER SCORING
(N = 20 Persons)

| Test Design | Correlations* of Criterion and Obtained Profile Scatter | |
|---|---|---|
| | Error-Free Data | Error Data |
| Unforced Test | .90 | .80 |
| Total Q-sort | .92 | .76 |
| Block Q-sort | .50<br>.72** | .64 |

*Relationships here reported are rank order correlations.

**This correlation was computed after eliminating one person from consideration. See text for explanation.

magnitude of this effect for the more operational test conditions represented by the rate of error used in this study.

The error-free correlations will be considered first. The $D_5$ similarity scores of the unforced test were computed between cluster profiles. Since the error-free $D_5$ scores were highly efficient in reproducing the total criterion similarity scores, and permitted unrestricted individual

scatter, the correlation of .90 between the criterion profile scatter and the cluster profile scatter for each person would be expected. The error $D_5$ scores, also, were relatively efficient in reproducing the $C_5$ scores, hence, the comparable relationship of .80 under error conditions is also consistent with the prior data for the unforced test.

The correlation of .92 between individual scatters within criterion profiles and individual scatters within cluster profiles from the error-free total Q-sort, however, is a different matter. This high relationship indicates that most of the scatter that results from cluster scoring closely approximates the original true scatter for each of the twenty subjects. The comparable correlation of .76 for the total Q-sort under the influence of error is an even stronger argument for the use of cluster scoring to measure differences in scatter that would not be treated by the Q or $DQ_3$ similarity scores.

Correlations given in Table 6 for block Q-sorts will be discussed in a later section when the efficiency of the block Q-sort is being considered.

The efficiency of the $DQ_4$ similarity scores. The correlations of the criterion score-sets and the $DQ_4$ score-sets for the total Q-sort will now be considered. The error-free $DQ_4$ correlations, as listed in Table 4 are $C_5$-$DQ_4$ = .68, $C_4$-$DQ_4$ = .74, and $C_3$-$DQ_4$ = .71. As would be expected from the preceding discussion concerning the value of cluster scoring, the $DQ_4$ scores have a slightly higher correlation with the $C_4$ scores than either the $C_5$ or $C_3$ scores. These three criterion-$DQ_4$ correlations are so nearly equivalent, however, that chance could account for any specific difference. These data indicate that the $DQ_4$ scores include information due to criterion differences in elevation and scatter. The $DQ_4$ scores, therefore, have some properties that are similar to the $DQ_3$ scores for the total Q-sort.

It will be recalled that the $DQ_4$ scores were computed between cluster profiles that had constant means over all persons but unrestricted individual scatter within profiles. These $DQ_4$ scores represent similarity scores computed between persons' positions located in a k - 1 hyperplane. Likewise, the $C_4$ scores represent similarity distances in a k - 1 hyperplane. As has been discussed, these two k - 1 hyperplanes will be identical, if and only if, the factor structure of the item sample is equivalent to the factor structure of the criterion space.

Since the factor weights of the item sample are not equivalent to the factor weights in the criterion space, the $DQ_4$ scores of this study are computed in a k - 1 hyperplane that is oblique to the corresponding criterion

k - 1 hyperplane. Thus, the information due to elevation that is eliminated in the $C_4$ score-set is not entirely eliminated in the $DQ_4$ score-set. This accounts for the fact that the $DQ_4$ score-set correlates approximately the same with all three criterion score-sets. In other words, by changing the weights of the factor loadings in the item sample the various criterion - $DQ_4$ correlations could be altered. As the item sample weights become more equivalent to the criterion weights, the $C_4$-$DQ_4$ correlation should increase and the $C_5$-$DQ_4$ and $C_3$-$DQ_4$ correlation should decrease since the criterion and obtained elevation dimension will tend to become more similar.

A comparison of the error-free and error relationships of the $DQ_4$ similarity scores for the total Q-sort provides further evidence of the value of the cluster scoring technique. The $DQ_4$ correlations under the influence of error are $C_5$-$DQ_4$ = .65, $C_4$-$DQ_4$ = .66, and $C_3$-$DQ_4$ = .38. Cluster scoring apparently reduces the effect of error since the $C_4$-$DQ_4$ and $C_5$-$DQ_4$ relationships remain relatively constant after the introduction of error. The nature of the clustering process suggests an explanation. The introduction of error into the system will tend to cause the Q-scores of some items within any cluster to increase and some Q scores to decrease. The summing process involved in obtaining each cluster score tends to nullify much of the effect of error since the sum of random errors over many items approaches zero. Consequently, each individual's cluster profiles are considerably alike under the presence and absence of error. These data suggest that cluster scoring methods may be of considerable value in improving the efficiency of Q-sort designs under operational error conditions.

Further studies will be necessary before the clustering effect can be more adequately quantified. The implications of the data of this study, however, strongly support the use of cluster scoring in situations wherein the test design is comparable to the total Q-sort as here described and the item sample contains homogeneous clusters.

The efficiency of the $D_3$ similarity scores for the total Q-sort. The five-score cluster profiles from which the $DQ_4$ similarity scores were computed have been described above. Since the cluster profiles have unrestricted five-score variances, it is possible to standardize these profiles by eliminating the individual differences in this variance. The $D_3$ similarity scores for the total Q-sort represent similarity distances between pairs of such standardized cluster profiles. The error-free correlations between the $D_3$ scores and the criterion scores as listed in Table 4 are as follows: $C_5$-$D_3$ = .36, $C_4$-$D_3$ = .40, and $C_3$-$D_3$ = .88. These data indicate that as expected the $D_3$ measures contain little information due to

criterion differences in elevation and scatter. The error-free $D_3$ score-set, however, is quite efficient for reproducing the criterion information due to profile differences in shape.

As error is added, the efficiency of the $D_3$ scores for reproducing the $C_3$ scores is decidedly diminished. The $D_3$ correlations under the influence of error are $C_5$-$D_3$ = .24, $C_4$-$D_3$ = .28, and $C_3$-$D_3$ = .38. As discussed above, the reduction of the information contained in the similarity measures by eliminating individual differences in scatter results in a magnification of the effect of error and, consequently, a reduction in the criterion-$D_3$ correlation.

Empirical evidence of the effect of error on the total Q-sort similarity scores. As discussed for the unforced test design, the effect of error can be empirically examined by comparing non-error similarity scores with corresponding similarity scores computed under the error condition. These relationships for the three obtained configurations of similarity scores for the total Q-sort are as follows: .45 for the $DQ_3$ scores, .70 for the $DQ_4$ scores, and .46 for the $D_3$ scores.

It is interesting to note that the effect of error on $DQ_3$ and $D_3$ scores was practically identical. It has been shown that the $DQ_3$ scores have eliminated scatter differences in the sixty-score profile with differences due to the five-score scatter being further eliminated in the $D_3$ scores. These data, however, suggest that the aforementioned magnification of error upon reduction to the k - 2 similarity measures occurs regardless of the amount of scatter information that has been eliminated. Of course, the above equality of correlation may be specific to this study; hence, further research is necessary to verify this suggestion.

The correlation of .70 between error-free and error $DQ_4$ scores indicates that the similarity scores computed between cluster profiles are relatively stable under the influence of error. This tendency reinforces the argument presented earlier that the summing process involved in cluster scoring tended to nullify the error effect. The $DQ_4$ scores, thus, appear to be of moderately high reliability and validity for measuring $C_4$ criterion scores.

General Conclusions Concerning the Use of the Total Q-sort Design in Similarity Studies.

The data of this study indicate that the total Q-sort can be moderately efficient for obtaining similarity scores under certain conditions. For perfectly reliable items the $D_3$ scores have high efficiency for reproducing

the $C_3$ criterion score-set. When error is introduced this efficiency shows a decided decrease.

The $DQ_3$ similarity scores are influenced by some specific conditions that are peculiar to this study. As a result, these scores have approximately equal correlations with all three criterion scores, and hence, are not highly efficient for reproducing any one set of criterion information.

Under the conditions of this study, some criterion information is apparently lost because of the forcing requirement of the Q-sort. As the amount of error in the system is increased, this information loss increases disproportionately.

The $DQ_4$ score-set seems to be the most promising for reproducing criterion similarity relationships from item samples that contain clusters of highly intercorrelated items. These $DQ_4$ scores have relatively high correlation with the $C_4$ and $C_5$ criterion scores under error-free conditions and also under the rate of error imposed in this study. It has been shown that certain characteristics of the factorial structure of this item sample tend to decrease the $C_4$-$DQ_4$ correlation and increase the $C_5$-$DQ_4$ relationship. Thus, as the factor structure of the item sample becomes more similar to the criterion factor structure, the efficiency of the $DQ_4$ scores for reproducing the $C_4$ criterion scores increases.

This same effect applies to the $DQ_3$ scores in that as the criterion and item factor weights become more similar, the $C_5$-$DQ_3$ correlation will tend to decrease and the $C_3$-$DQ_3$ correlation will tend to increase. In other words, the elimination or reduction of certain conditions specific to this study will cause an increase in the efficiency of the $DQ_3$ scores for reproducing the $C_3$ score-set under error-free conditions. However, the $DQ_3$ scores are subject to the effects of magnification of error and, as a result, the $C_3$-$DQ_3$ relationship will be seriously reduced under practical error conditions.

The $C_4$-$DQ_4$ relationships suggest that total Q-sort item samples should be structured so that cluster scoring is feasible. Operationally, this would involve a logical or empirical factor analysis of suggested items in order that homogeneous clusters could be established. The data of this study indicate that the additional work involved in establishing such an item sample would be well repaid by increased validities and reliabilities. Further research in this area should lead to a more adequate understanding of the effects of cluster scoring on similarity measures obtained under the total Q-sort design.

## The Efficiency of the Block Q-Sort Design

The data relating to the efficiency of the block Q-sort design as specified in this study will now be considered. Information losses due to this particular design will be discussed. The effect of cluster scoring will be considered. The effect of the range of item popularity within the block structure will be examined. General conclusions concerning the use of the block Q-sort design in similarity studies will be developed as a consequence of the relations indicated in this study.

A brief review of the structure of this block Q-sort may be helpful prior to the discussion of the criterion correlations for this design. Each block contains five items of equal popularity, with each item having a loading of .70 or higher on a different factor. In cluster scoring, therefore, each factor cluster contains one item from each of the twelve blocks.

Each block of five items is treated as a small, complete Q-sort. Each subject chooses the item most positive for him and the item most negative for him in each block. This design, therefore, is more comparable to a composite of twelve similar five-item Q-sorts, than to a sixty-item Q-sort. The efficiency of this design will now be examined.

### Test-Criterion Relationships

The correlations between the criterion and obtained similarity score-sets for the block Q-sort are presented in Table 7. The error-free correlations will be considered first. These data indicate that for perfectly reliable data the block Q-sort is relatively inflexible. All three obtained similarity score-sets have moderately high correlations with the $C_3$ score-set and comparatively low relationships with the $C_4$ and $C_5$ score-sets. These data indicate that the obtained scores of the block Q-sort have eliminated most of the criterion information due to differences in elevation and scatter. As a result, these obtained scores are quite efficient for reproducing criterion information due to profile shape differences under error-free conditions.

The criterion-obtained correlations of the block design reflect some of the specific conditions peculiar to this study that are discussed above. In particular, the $DQ_3$ and $DQ_4$ scores include some criterion elevation because of the unequal factor weights of the item sample. This tends to cause some increase in the correlation of these scores with the $C_5$ scores. Since the $DQ_3$ scores are computed between profiles that have thirty-six degrees of freedom, i.e., permit thirty-six independent judgments in specifying the profile, this score-set contains some scatter that

TABLE 7

CORRELATIONS OF CRITERION SCORE-SETS AND OBTAINED SIMILARITY SCORE-SETS FOR THE BLOCK Q-SORT DESIGN*

| Obtained Similarity Score-Sets | Perfectly Reliable Items | | | Moderately Reliable Items | | |
|---|---|---|---|---|---|---|
| | Criterion Score-Sets | | | Criterion Score-Sets | | |
| | $C_5$ | $C_4$ | $C_3$ | $C_5$ | $C_4$ | $C_3$ |
| $DQ_3$** | .47 | .42 | .78 | .12 | .10 | .28 |
| $DQ_4$ | .43 | .53 | .81 | .51 | .48 | .19 |
| $D_3$ | .36 | .39 | .84 | .20 | .20 | .42 |

* The correlation between each criterion score-set and its most logically related, i.e., most relevant, obtained score-set is underlined. Correlations that represent semi-relevant relationships between criterion and obtained score-sets due to specific conditions of this study are underlined with dots.

**The D measures for these configurations were computed from Q correlations.

is eliminated in the five-score profiles from which the $C_3$ score-set is computed. This causes a slight decrease in the $C_3$-$DQ_3$ correlation and would suggest that the $DQ_3$ scores would be even more efficient for reproducing the $C_3$ score-set if these specific conditions were eliminated or reduced.

The comparable correlations after error has been introduced however, are considerably less encouraging. As was expected, the relatively high $C_3$ correlations for the error-free case decrease to very low relationships due to the magnification of error in the k - 2 space. Only the $DQ_4$ correlations with the $C_4$ and $C_5$ criterion scores are of sufficient magnitude to warrant further discussion.

## The Effect of Cluster Scoring for the Block Q-sort

As has been discussed, the effect of cluster scoring can be empirically examined by correlating the criterion profile scatter of each individual with the corresponding cluster profile scatter obtained from the test design under consideration. These relationships for the test designs considered in this study were presented in Table 6 for the perfectly reliable and less reliable items.

The error-free correlation between the individual scatter within the criterion profiles and the individual scatter within the cluster profiles was .50 for the block Q-sort. This relationship was considerably lower than for the other test designs. In examining the computational processes, however, it was observed that this lower correlation was somewhat misleading since most of the reduced relationship was caused by an extreme shift in the rank of one person. In view of the unusual variation of this person's obtained profile scatter a second correlation was computed for nineteen persons with the atypical person eliminated. Under this condition the correlation between the criterion and cluster profile variances increased to .72. Just how many other undiscovered specific effects exist in this block Q-sort one cannot say, but in any case, the individual differences in variance within the five-score cluster profiles are moderately related to the criterion variances.

Under the influence of error the comparable correlation between criterion profile scatter and cluster profile scatter for the block Q-sort, as listed in Table 6, is .64. This supports the above inference that there is some value in cluster scoring in the block design. Since this relationship is considerably smaller than for either the unforced or total Q-sort tests under the same rate of error, however, it is further indication of the general inefficiency and inflexibility of the block design as compared with the other two experimental designs.

## The Characteristics of the $DQ_4$ Measures for the Block Q-Sort

In view of the above data indicating the value of cluster scoring in this block Q-sort design, further consideration of the $DQ_4$ similarity measures is necessary. The $C_5$-$DQ_4$ correlations are .43 and .51 and the $C_4$-$DQ_4$ correlations are .53 and .48 under error-free and error conditions, respectively. The relative stability of these relationships under the absence and presence of error suggests that the effect of error is reduced by the clustering process.

A further discussion of the characteristics of the block $DQ_4$ scoresets is necessary. The correlations of $C_5$ -$DQ_4$ = .51 and $C_4$-$DQ_4$ = .48 under the influence of error indicate that these $DQ_4$ scores contain some criterion information due to difference in elevation and scatter. Since the

item sample has not been altered, the criterion elevation component in the $DQ_4$ scores results from the unequal factor weightings in the factor space and the item sample as already discussed. Consequently, cluster scoring seems of some value in this block design and would probably be of more value if some of these specific effects could be reduced.

Some additional information concerning the treatment of error in cluster scoring should be given by the correlation between the error-free and error block $DQ_4$ score-sets. This correlation was computed as .28. Since the $DQ_4$ scores do have some validity for the specified rate of error, this low reliability estimate is somewhat perplexing and probably misleading. However, since the error and non-error correlations of the $DQ_3$ and $D_3$ scores are also low, .34 and .45, respectively, the reliability of any block Q-sort similarity score is likely to be relatively low for any operational rate of error.[10]

General Conclusions Concerning the Use of the Block Q-Sort Design in Similarity Studies

The data of this study indicate that the block design as here described is a relatively inflexible and inefficient type of instrument for measuring the similarity between persons under practical conditions. If the items are highly reliable the k - 2 criterion scores can be reproduced with considerable accuracy. The $DQ_4$ scores computed after cluster scoring offer some possibility for reproducing the $C_5$ and $C_4$ criterion scores but the indicated relationships are rather low.

Apparently, the disadvantages of the Q-sort are magnified and multiplied by the small subtest Q-sort design. In particular, small idiosyncratic tendencies that appear in each subtest assume large proportions when combined several times. This demonstrates the need for further theory relating to the forced-choice design test, in particular, those types that are composites of many small forced-choice subtests.

---

[10] This inference is supported by a study by Rudin and others (14). They investigated the reliability of a block Q-sort type instrument comparable to the one under discussion. Their conclusion as to the reliability of their block Q-sort instrument for measuring the similarity between persons was as follows: "Measures of Real Similarity on the present instrument were very unreliable and cannot be expected to correlate with criteria. The reliability of these measures increased slightly but not sufficiently by item selection. Profile scoring was of no help. However, it should be noted that the profile scoring procedure was not finally tested here, having been limited to five-item blocks and clusters of low internal consistency." (14, pp. 11)

## Comparison of the Efficiency of the Three Experimental Test Designs

The correlations of the criterion score-sets and the obtained similarity score-sets for the three experimental test designs are compared in Table 8. Since each criterion score-set should be reproduced most highly by those obtained score-sets that include comparable information, the correlations between criterion and logically related obtained score-sets are underlined. The largest correlation between each criterion score-set and a logically related obtained score-set is doubly underlined for both error-free and error conditions. Those semi-relevant criterion-obtained correlations, i.e., relationships between criterion and obtained score-sets that are logical because of specific conditions of this study, are indicated by underlining with dots. Inspection of these data indicates certain trends or patterns of relationships between the criterion scores and the obtained scores for different test designs.

For example, in every case, i.e., for all criterion score-sets under both the presence and absence of error, the largest efficiency correlation is produced by an obtained score-set of the unforced test design. These data indicate that under the conditions specified in this study the unforced test design has some superiority in both efficiency and flexibility over the two experimental Q-sort designs.

This superiority of the unforced design is greatest for reproducing the $C_5$ score-set and reduces to practical insignificance for the $C_3$ scores.

As discussed earlier, these relationships for the unforced test must be interpreted with caution since this experimental unforced design will be difficult to reproduce in practice. These data, however, do suggest that the unforced design has the greatest potential value for measuring similarity between persons. The degree to which this potential value is realized depends upon the extent to which response sets can be eliminated, or included if valid measures, and the clarity with which the multi-scale positions can be presented.

Further inspection of the correlations of Table 8 indicates that the total Q-sort tends to be more efficient for reproducing the $C_4$ score-set than the block Q-sort. While some of the differences between these relationships are probably due to chance effects, the overall trends suggests some slight superiority for the total Q-sort.

TABLE 8

CORRELATIONS OF CRITERION SCORE-SETS AND OBTAINED SIMILARITY SCORE-SETS FOR THE THREE EXPERIMENTAL TEST DESIGNS*

| Type of Obtained Score-Set | Criterion Score-Set | Perfectly Reliable Items | | | | | Moderately Reliable Items | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obtained Similarity Score-Sets | | | | | Obtained Similarity Score-Sets | | | | |
| | | $D_5$ | $D_4$ | $DQ_3$ | $DQ_4$ | $D_3$ | $D_5$ | $D_4$ | $DQ_3$ | $DQ_4$ | $D_3$ |
| Unforced Design | $C_5$ | .92 | .85 | | | .39 | .81 | .71 | | | .26 |
| Total Q-sort | | | | .78 | .68 | .36 | | | .22 | .65 | .24 |
| Block Q-sort | | | | .47 | .43 | .36 | | | .12 | .51 | .20 |
| Unforced Design | $C_4$ | .58 | .93 | | | .44 | .55 | .77 | | | .31 |
| Total Q-sort | | | | .70 | .74 | .40 | | | .18 | .66 | .28 |
| Block Q-sort | | | | .40 | .53 | .39 | | | .10 | .44 | .20 |
| Unforced Design | $C_3$ | .27 | .46 | | | .90 | .25 | .31 | | | .45 |
| Total Q-sort | | | | .66 | .71 | .88 | | | .38 | .38 | .38 |
| Block Q-sort | | | | .78 | .81 | .84 | | | .28 | .19 | .42 |

*The correlation between each criterion score-set and its most relevant obtained score-set is singly underlined; the largest such relationship for each criterion score-set is doubly underlined. Semi-relevant relationships due to the specific conditions of this study are underlined with dots.

## The Implications of the $C_3$ Correlations

The matrix containing the correlations of $C_3$ scores with the various obtained similarity scores under the rate of error specified in this study is of considerable interest. In particular, note the generally low magnitudes and limited range of correlations for the most related obtained score-sets, i.e., from .45 to .28. These data indicate that with this rate of error all three test types are uniformily inefficient in reproducing this criterion configuration of similarity scores. The maximum $C_3$ correlation of .45 indicates that approximately 20 percent of the $C_3$ information is being reproduced by this obtained configuration of similarity scores. Considering that the $C_3$ criterion scores contain approximately 15 percent of the total $C_5$ criterion information, one realizes just how small a portion of the total criterion variance is being reproduced by obtained measures that operate in k - 2 dimensions.

These data indicate that $C_3$ information is difficult to measure accurately under practical rates of error. The general implications of this study have suggested the inadvisability of using similarity scores that ignore all individual differences in elevation and scatter. Some test situations may arise, however, in which elevation information is completely irrelevant and information due to scatter can be eliminated on theoretical or empirical grounds. In such a case, the obtained similarity scores should operate in k - 2 dimensions and should not measure individual differences in elevation and scatter. If the test conditions are similar to those of this study, however, it will be necessary to use an extremely long test or to construct highly reliable items in order to accurately reproduce the desired $C_3$ criterion information.

## Limitations of this Study

It is important that the restricted nature of this test situation be kept in mind in interpreting the results of this study. Furthermore, it must be remembered that only twenty persons are included in the sample of persons under consideration. Thus, some sampling effects will be included in the computed data that are specific to this study. Consequently, statements as to the superiority of one test design over another cannot be generalized to other test situations from this study alone.

This study, however, tends to support the position that properly structured unforced designs will be most valid for measuring similarity between persons. Many limiting assumptions concerning the number of factors, distribution of errors, assignment of obtained scores, etc., were made of necessity to keep this investigation operationally feasible. As a consequence, the implications of the empirical data presented here must be interpreted with caution. The major contribution of this study is seen as opening up a relatively new methodological approach to the study of

similarity measures. This claim seems justified since the theoretical model and computational techniques developed in this study and available elsewhere (19) are quite general and can be used to further investigate many of the unanswered problems that have arisen due to the limitations of these test conditions. Therefore, most of the implications of this investigation should be valuable in the structuring of further research and must not be over-generalized as to their immediate application.

One important specific limitation of this item structure that has implications for generalizing the efficiency of the total Q-sort to other studies should be discussed briefly. This limitation concerns the a priori specification of a positive direction for all factors and the restriction that all item vectors have this same positive direction (see page 7). Our present understanding of the problem is not sufficient to permit us to predict how the total Q-sort structured under these conditions would differ from a total Q-sort structured under the conditions originally specified by Stephenson (17). He suggested that items should be selected for the Q-sort sample in pairs, such that for every positively oriented item vector there should be a comparable item vector orientated in the opposite direction. For example, if a Q-sort item sample contains an item positively loaded on extroversion there should be a comparable item positively loaded on introversion. Since the results of this study were obtained under a specific design, additional studies must be completed before an exhaustive answer can be given as to the efficiency of the total Q-sort as a general technique for measuring similarities between persons.

## Conclusions and Recommendations

With these limitations in mind, the following conclusions can be discussed as developing from this study.

1. Under the conditions of this study, the experimental unforced test design is more efficient that the two experimental Q-sort test designs for measuring the similarity between persons.

   a) Advantages of the unforced test design are,

      (1) All criterion score-sets are highly reproduced by logically related unforced score-sets under error-free conditions.

      (2) The k and k - 1 criterion score-sets are highly reproduced by logically related unforced score-sets under the rate of error introduced in this study.

      (3) Each unforced score-set has higher efficiency for reproducing the related criterion score-set under the error conditions of this study than the comparable score-set of either of the Q-sort designs.

b) Disadvantages of the unforced test are,

(1) It is subject to undesirable response-sets,

(a) These response-sets can be reduced by carefully structuring the test items and the test instructions.

(2) It is difficult to construct an unforced test that will give unforced, continuous responses.

(a) A clearly defined multi-position scale for each item is probably a reasonable approximation.

(1) Further research is needed to compare the efficiencies of an unforced test that permits continuous scoring with an unforced test that requires discontinuous scoring and, hence, permits tied scores.

2. Some criterion information is lost due to the forcing requirement when Q-sort designs are used to measure the similarity between persons.

a) Test designs requiring smaller Q-sorts within the larger item sample, e.g., the block design, tend to increase this information loss by magnifying differences due to error or individual idiosycrasies.

b) The efficiency of the scores from the Q-sort design decreases as the factor structure of the item sample deviates from the factor structure of the specified criterion

(1) When the factor weights of the item sample differ from those of criterion space, the obtained test scores include some criterion information due to individual differences in profile elevation, scatter, and shape. Under such conditions the Q-sort scores are not efficient for reproducing any one criterion score-set.

c) Cluster scoring tends to increase the efficiency of the Q-sort test design for reproducing the criterion score-set that includes information due to differences in profile scatter and shape.

(1) This suggests that Q-sort item samples should be designed to contain relatively independent clusters of homogeneous items in order to permit cluster scoring.

d) Criterion scores that have eliminated information due to individual differences in elevation and scatter can be highly reproduced by the most related scores of the Q-sort design for highly reliable items.

(1) When error is introduced into the system, this efficiency shows a decided decrease.

3. The reduction of the dimensions of similarity measures by eliminating individual differences in elevation, or elevation and scatter, has extremely important implications for studies of similarity between persons.

a) Under the conditions of this study, a considerable portion of the total criterion information is lost in such reductions.

(1) When elevation is eliminated, 33 percent of the total criterion information is lost.

(2) When elevation and scatter are eliminated, 85 percent of the total criterion information is lost.

b) The reduction of the dimensions of similarity scores from k - 1 to k - 2 by eliminating differences in scatter causes a definite magnification of the effect of any error present in the system.

(1) As a result, similarity scores that eliminate differences in scatter are generally inefficient under practical error conditions regardless of the test design from which they were obtained.

These conclusions have many implications for researchers and practical testers who are interested in measuring the similarities between persons in a test structure as described in this study. While many of the above conclusions need additional support before wide generalization is possible, these recommendations seem reasonable at this time.

1. The results of this study suggest that similarity measures should be obtained using a carefully constructed unforced test design since this type seems to be the most efficient for measuring similarity relationships between persons.

a) Considerable care should be taken to avoid irrelevant response-sets in the unforced test, but for some similarity scores certain relevant response-sets may actually contribute to the validity of the measure.

2. If forced-choice test designs are used to measure similarity relationships, the greatest amount of criterion information can be consistently reproduced by cluster scoring the homogeneous items before computing similarity scores.

a) The raw item scores should not be used to estimate similarity relationships unless the factor weights of the item sample have been carefully selected to be consistent with the factor weights of the criterion measures.

(1) The use of such inefficient item scores, e.g., scores from inadequate item samples and/or from small forced-choice block arrangements, will consistently tend to obscure true relationships and may yield insignificant results where significant differences actually exist.

3. Criterion similarity relationships that ignore differences in elevation and scatter are extremely difficult to reproduce by any of the test designs used in this study.

a) When theory or empirical evidence requires the measurement of such relationships, it will probably be necessary to use very long tests or to construct highly reliable items if efficient measurement is to be accomplished.

## BIBLIOGRAPHY


1. Burt, C. Correlation between persons. Brit. J. Psychol., 1937, 28, 59-95.

2. Burt, C. The factors of the mind. London: Univ. of London Press, 1940.

3. Cronbach, L. J. Essentials of psychological testing. New York: Harper, 1949.

4. Cronbach, L. J. Further evidence on response-sets and test design. Educ. psychol. Measmt., 1950, 10, 3-31.

5. Cronbach, J. J., and Gleser, Goldine C. Similarity between persons and problems of profile analysis. Champaign-Urbana, Illinois: Univ. of Illinois, 1952. (Mimeographed, Technical Report No. 2, Contract N6ori-07135 between the University of Illinois and the Office of Naval Research).

6. Eberman, P. W. Personal relationships: one key to instructional improvement. Educ. Leadership, 1952, 9, 389-392.

7. Edelson, M., and Jones, A. E. The use of Q-technique and role-playing in an investigation of the self-concept. Unpublished manuscript, Univ. of Chicago, 1951. (private circulation).

8. Fiedler, F. E. A comparison of therapeutic relationships in psychoanalytic, non-directive, and Adlerian therapy. J. consult. Psychol., 1950, 14, 436-446.

9. Fiedler, F. E., Hartmann, W., and Rudin, S. A. The relationship of interpersonal perception to effectiveness in basketball teams. Champaign-Urbana, Illinois: Univ. of Illinois, 1952. (Mimeographed Technical Report No. 3, Contract N6ori-07135 between the University of Illinois and the Office of Naval Research).

10. Fiedler, F. E., Blaisdell, F. J., and Warrington, W. G. Unconscious attitudes and the dynamics of sociometric choice in a social group. J. abnorm. soc. Psychol., 1952, 47, 790-796.

11. Gordon, L. V. Validities of the forced-choice and questionnaire methods of personality measurement. J. appl. Psychol., 1951, 35, 407-412.

12. McNemar, Q. Psychological statistics. New York: Wiley, 1949.

13. Osgood, C. E., and Suci, G. J. A measure of the relation determined by both mean difference and profile information. Psychol. Bull. 1952, 49, 251-262.

14. Rudin, S. A., Lazar, I., Ehart, Mary E., and Cronbach, L. J. Some empirical studies of the reliability of social perception scores. Champaign-Urbana, Illinois: Univ. of Illinois, 1952. (Mimeographed Technical Report No. 4, Contract N6ori-07135 between University of Illinois and the Office of Naval Research.)

15. Stephenson, W. Correlating persons instead of tests. Character and Pers., 1935, 4, 17-24.

16. Stephenson, W. Methodological consideration of Jung's typology. J. ment. Sci., 1939, 85, 185-205.

17. Stephenson, W. Q-technique: variate-designs and propositional sets. Unpublished monograph, Univ. of Chicago, 1951.

18. Stephenson, W. A note on Professor R. B. Cattell's methodological adumbrations. J. clin. Psychol., 1952, 8, 206-207.

19. Warrington, W. G. The efficiency of the Q-sort and other test designs for measuring the similarity between persons. Unpublished Doctoral Thesis. University of Illinois. 1952.

20. Zuckerman, J. V. Interest item response arrangement as it affects discrimination between professional groups. J. appl. Psychol., 1952, 36, 79-85.